An iterative hard thresholding algorithm with improved convergence for low-rank tensor recovery

José Henrique de Morais Goulart, Gérard Favier

AN ITERATIVE HARD THRESHOLDING ALGORITHM WITH IMPROVED CONVERGENCE FOR LOW-RANK TENSOR RECOVERY

José Henrique de M. Goulart,* *Gérard Favier*

I3S Laboratory, Université de Nice Sophia Antipolis, CNRS, France

ABSTRACT

Recovering low-rank tensors from undercomplete linear measurements is a computationally challenging problem of great practical importance. Most existing approaches circumvent the intractability of the tensor rank by considering instead the multilinear rank. Among them, the recently proposed tensor iterative hard thresholding (TIHT) algorithm is simple and has low cost per iteration, but converges quite slowly. In this work, we propose a new step size selection heuristic for accelerating its convergence, relying on a condition which (ideally) ensures monotonic decrease of its target cost function. This condition is obtained by studying TIHT from the standpoint of the majorization-minimization strategy which underlies the normalized IHT algorithm used for sparse vector recovery. Simulation results are presented for synthetic data tensor recovery and brain MRI data tensor completion, showing that the performance of TIHT is notably improved by our heuristic, with a small to moderate increase of the cost per iteration.

***Index Terms*—** Low-rank Tensor Recovery, Tensor Completion, Iterative Hard Thresholding

1. INTRODUCTION

Tensors having (approximately) low rank arise in many practical applications. Whenever true, this property can in principle be exploited to recover a tensor of interest from *undercomplete information* given by *linear* observations, a task which is ill-posed in general. An important special case of this setting is the completion of a data tensor having missing entries under the low-rank assumption. These problems, called low-rank tensor recovery (LRTR) and tensor completion (TC), respectively, are extensions of low-rank matrix recovery (LRMR) and matrix completion [1], and find several applications such as image inpainting [2], seismic signal processing [3], spectral data recovery [4] and machine learning [5].

However, despite being a natural generalization of the matrix rank, the tensor rank is not completely understood and is computationally intractable [6]. Consequently, many existing LRTR techniques rely instead on the multilinear rank, which is a multi-valued quantity composed by the ranks of all mode-n matrix unfoldings [7,8]. This choice is motivated by the fact that the tensor rank upper bounds the rank of each unfolding.

Among the major approaches, a common one is to seek the joint minimization of the nuclear norms (NN) of the mode-n unfoldings, instead of their ranks. Its popularity stems from the effectiveness of the NN minimization (NNM) approach for LRMR [1, 9]. In the tensor case, one usually introduces a regularization functional given by a weighted sum of the nuclear norms of these unfoldings [2, 4, 10]. Yet, [9] shows that this cannot be more efficient, in terms of the minimal number of measurements needed, than solely minimizing the NN of the "best" unfolding in that sense, which is quite far from the theoretical optimal [11]. Although the NNM of a more "balanced" matrix unfolding can get closer to the optimal number of necessary measurements [11], this applies only to tensors of order $P > 3$, and a significant gap still remains. Another possibility consists in directly estimating a low-rank tensor model via alternating minimization of a data-based error criterion, as in [5]. This performs often well in practice, but can be quite difficult to analyze—to the best of our knowledge, no global convergence proofs are known for the alternating estimation of standard low-rank models, unless additional regularization is used [12].

Recently, [13] has proposed a simple and effective algorithm called tensor iterative hard thresholding (TIHT). Basically, it can be seen as a multilinear-rank-based variant of the normalized IHT (NIHT) algorithm proposed in [14] for sparse vector recovery. Even though no convergence proofs and performance bounds exist yet for TIHT, it is simple to implement and is less costly than the above mentioned approaches. However, its convergence speed observed in numerical experiments is quite slow.

In this paper, we study the TIHT algorithm from the standpoint of the majorization-minimization (MM) strategy proposed in [14]. This enables us to obtain an upper bound for the step size which guarantees that the iterates have monotonically decreasing cost function values in the ideal case where the best low-rank approximation computed at each iteration is exact. Then, by exploiting this bound, we propose an algorithm named improved-step-selection TIHT (ISS-TIHT) containing a heuristic subroutine which attempts to find a step size within a constant factor of its upper bound. Our simulation results show that this remarkably improves conver-

* Sponsored by CNPq–Brazil (individual grant 245358/2012-9).

gence speed, which significantly compensates for the increase in computational cost.

2. TENSOR ITERATIVE HARD THRESHOLDING

Let $\boldsymbol{\mathcal{T}} \in U = \mathbb{R}^{N_1 \times \cdots \times N_P}$ be a Pth-order tensor and denote by $\mathbf{T}_p$ the mode-p unfolding (matricization) of $\boldsymbol{\mathcal{T}}$ [7]. The *multilinear rank* (m-rank) of $\boldsymbol{\mathcal{T}}$ is a P-tuple given by $\text{m-rank}(\boldsymbol{\mathcal{T}}) = (\text{rank}(\mathbf{T}_1), \ldots, \text{rank}(\mathbf{T}_P))$ [8].

In view of the computational difficulty of minimizing the tensor rank, the LRTR problem is tackled in [13] by imposing a component-wise bound $\mathbf{r} = (R_1, \ldots, R_P)$ on the m-rank of the solution, which is sought in the least-squares sense in $L_{\mathbf{r}} = \{\boldsymbol{\mathcal{T}} \in U : \text{rank}(\mathbf{T}_p) \leq R_p,\ p = 1, \ldots, P\}$. This leads to the constrained formulation

$$\min_{\boldsymbol{\mathcal{T}} \in L_{\mathbf{r}}} J(\boldsymbol{\mathcal{T}}), \quad \text{with} \quad J(\boldsymbol{\mathcal{T}}) = \|\mathbf{y} - \mathscr{A}(\boldsymbol{\mathcal{T}})\|_2^2, \qquad (1)$$

where $\mathscr{A} : U \mapsto \mathbb{R}^M$ is a linear measurement operator and $\mathbf{y} \in \mathbb{R}^M$ is a given vector of measurements. We denote by $\mathbf{A}$ the matrix representation of $\mathscr{A}$ such that $\mathscr{A}(\boldsymbol{\mathcal{T}}) = \mathbf{A}\,\text{vec}(\boldsymbol{\mathcal{T}})$, where $\text{vec}(\cdot)$ stacks the elements of its argument in a long vector. Note that (1) applies in particular to the TC problem, where $\mathscr{A}$ is constrained to be a sampling operator; in other words, $\mathbf{A}$ has M canonical vectors of $\mathbb{R}^{N_1 \cdots N_P}$ as rows.

To solve (1), [13] proposes the TIHT algorithm

$$\boldsymbol{\mathcal{T}}_k = \mathscr{H}_{\mathbf{r}}\left(\boldsymbol{\mathcal{T}}_{k-1} + \mu_k \mathscr{A}^*\left(\mathbf{y} - \mathscr{A}\left(\boldsymbol{\mathcal{T}}_{k-1}\right)\right)\right), \qquad (2)$$

where $\mathscr{A}^* : \mathbb{R}^M \mapsto U$ is the adjoint of $\mathscr{A}$ (*i.e.*, $\mathscr{A}^*(\mathbf{x}) = \text{unvec}(\mathbf{A}^T\mathbf{x})$, where $\text{unvec}(\cdot)$ is the inverse of $\text{vec}(\cdot)$), $\mu_k > 0$ is a step size parameter and $\mathscr{H}_{\mathbf{r}} : U \mapsto L_{\mathbf{r}}$ maps a tensor $\boldsymbol{\mathcal{T}}$ into the $\mathbf{r}$-truncation of its higher-order SVD (HOSVD). More concretely, writing this HOSVD as $\boldsymbol{\mathcal{T}} = \boldsymbol{\mathcal{S}} \times_1 \mathbf{U}^{(1)} \times_2 \cdots \times_P \mathbf{U}^{(P)}$, where $\boldsymbol{\mathcal{S}} \in U$ is the core tensor and $\mathbf{U}^{(p)} \in \mathbb{R}^{N_p \times N_p}$ is the matrix of pth-mode singular vectors, $\mathscr{H}_{\mathbf{r}}(\boldsymbol{\mathcal{T}}) = \bar{\boldsymbol{\mathcal{S}}} \times_1 \bar{\mathbf{U}}^{(1)} \cdots \times_P \bar{\mathbf{U}}^{(P)}$, where $\bar{\mathbf{U}}^{(p)}$ contains the first R_p columns of $\mathbf{U}^{(p)}$ and $\bar{\boldsymbol{\mathcal{S}}} \in \mathbb{R}^{R_1 \times \cdots \times R_P}$ satisfies $[\bar{\boldsymbol{\mathcal{S}}}]_{r_1,\ldots,r_P} = [\boldsymbol{\mathcal{S}}]_{r_1,\ldots,r_P}$ for all $r_p \in \{1, \ldots, R_p\}$. The formula given in [13] for μ_k can be written as

$$\mu_k = \|\boldsymbol{\mathcal{G}}_{k-1}\|_F^2 \, \|\mathscr{A}(\boldsymbol{\mathcal{G}}_{k-1})\|_2^{-2}, \qquad (3)$$

where $\boldsymbol{\mathcal{G}}_{k-1} \triangleq -\mathscr{A}^*\left(\mathbf{y} - \mathscr{A}\left(\boldsymbol{\mathcal{T}}_{k-1}\right)\right)$. It is easy to show that, in the TC setting, $\mu_k = 1$ and the TIHT algorithm is equivalent to the HOSVD-based scheme proposed in [3] (with $a = 1$), because of the special form of $\mathscr{A}$.

As $2\boldsymbol{\mathcal{G}}_{k-1}$ is the gradient of J at $\boldsymbol{\mathcal{T}}_{k-1}$, one can see that TIHT first updates the current estimate $\boldsymbol{\mathcal{T}}_{k-1}$ with a gradient descent step and then computes an approximation of the result in the feasible set. This is therefore very similar in spirit to the projected gradient algorithm [15], in which such approximation is the projection onto the feasible set. Yet, as projecting onto $L_{\mathbf{r}}$ amounts to solving a best rank-$(R_1, \ldots, R_P)$ approximation problem, which is NP-Hard [6] and requires using costly algorithms (see, *e.g.*, [16]), a low-rank approximation given by the truncated HOSVD is used instead. This is a widely used technique which, despite being suboptimal, gives an approximant satisfying $\|\mathscr{H}_{\mathbf{r}}(\boldsymbol{\mathcal{T}}) - \boldsymbol{\mathcal{T}}\|_F \leq \sqrt{P}\|\boldsymbol{\mathcal{T}}^b - \boldsymbol{\mathcal{T}}\|_F$, where $\boldsymbol{\mathcal{T}}^b$ is a minimizer of the Euclidian distance to $\boldsymbol{\mathcal{T}}$ in $L_{\mathbf{r}}$ [8]. Apart from the suboptimality of $\mathscr{H}_{\mathbf{r}}$, it is important to point out that the optimality condition underlying the projected gradient algorithm applies only to *convex* feasible sets [15], which is not the case for $L_{\mathbf{r}}$. One might then wonder how TIHT actually achieves recovery.

In the following, relying on the optimization strategy developed by [14], we further study the TIHT algorithm. This study will then serve as a basis for devising a new step size selection routine in order to improve its convergence speed.

3. MONOTONICALLY DECREASING OBJECTIVE VALUES VIA MAJORIZATION-MINIMIZATION

The NIHT algorithm proposed in [14] is based on a clever MM strategy devised to minimize $\|\mathbf{y} - \boldsymbol{\Phi}\mathbf{x}\|_2^2$ subject to $\mathbf{x} \in V_s \subset \mathbb{R}^N$, where V_s is the subset of s-sparse vectors of $\mathbb{R}^N$ and $\mathbf{y} \in \mathbb{R}^M$ is the linear measurement of an s-sparse vector of interest given by $\boldsymbol{\Phi} \in \mathbb{R}^{M \times N}$. As shown in [14], the NIHT iterates have monotonically decreasing cost function (or objective) values and are convergent. Yet, whether this strategy promptly carries over to other similar problems, as (1), is not immediately clear. In what follows, we show that this is true for problem (1) and that the TIHT algorithm can be interpreted as an extension of this MM approach.

Recall that, to minimize a cost function J, an MM algorithm proceeds by minimizing instead at each iteration k a surrogate function J_k which majorizes J and coincides with it at the current estimate. It is not difficult to see that iterates computed in that manner are driven downhill with respect to J. The interest lies in the possibility of constructing surrogate functions J_k which are easier to minimize than J under the considered constraints.

In the case of TIHT, given the current estimate $\boldsymbol{\mathcal{T}}_{k-1}$ and a constant μ_k such that

$$J_k(\boldsymbol{\mathcal{T}}) \triangleq \mu_k J(\boldsymbol{\mathcal{T}}) + \|\boldsymbol{\mathcal{T}} - \boldsymbol{\mathcal{T}}_{k-1}\|_F^2 - \mu_k \|\mathscr{A}(\boldsymbol{\mathcal{T}} - \boldsymbol{\mathcal{T}}_{k-1})\|_2^2 \qquad (4)$$

satisfies $\mu_k J(\boldsymbol{\mathcal{T}}) < J_k(\boldsymbol{\mathcal{T}})$ for all $\boldsymbol{\mathcal{T}} \neq \boldsymbol{\mathcal{T}}_{k-1}$, we minimize J_k over $L_{\mathbf{r}}$ to obtain a new estimate $\boldsymbol{\mathcal{T}}_k$. Note that such a μ_k always exists, since we can choose it such that $0 < \mu_k < \|\mathscr{A}\|^{-2}$. Hence, if the minimizer $\boldsymbol{\mathcal{T}}_k$ of J_k over $L_{\mathbf{r}}$ satisfies $\boldsymbol{\mathcal{T}}_k \neq \boldsymbol{\mathcal{T}}_{k-1}$, then we have $\mu_k J(\boldsymbol{\mathcal{T}}_k) < J_k(\boldsymbol{\mathcal{T}}_k) \leq J_k(\boldsymbol{\mathcal{T}}_{k-1}) = \mu_k J(\boldsymbol{\mathcal{T}}_{k-1})$, thus yielding $J(\boldsymbol{\mathcal{T}}_k) < J(\boldsymbol{\mathcal{T}}_{k-1})$.

Let us now consider the minimization of J_k over $L_{\mathbf{r}}$. Replacing $J(\boldsymbol{\mathcal{T}})$ by (1) in (4), we have $J_k(\boldsymbol{\mathcal{T}}) = \|\boldsymbol{\mathcal{T}} - \boldsymbol{\mathcal{T}}_{k-1}\|_F^2 + \mu_k\|\mathbf{y}\|_2^2 - 2\mu_k\langle \mathbf{y} - \mathscr{A}(\boldsymbol{\mathcal{T}}_{k-1}), \mathscr{A}(\boldsymbol{\mathcal{T}})\rangle - \mu_k\|\mathscr{A}(\boldsymbol{\mathcal{T}}_{k-1})\|_2^2$, which is clearly strictly convex. Therefore, solving $J_k'(\boldsymbol{\mathcal{T}}^\star) = 0$ yields the unique unconstrained minimizer

$$\boldsymbol{\mathcal{T}}^\star = \boldsymbol{\mathcal{T}}_{k-1} + \mu_k \mathscr{A}^*(\mathbf{y} - \mathscr{A}(\boldsymbol{\mathcal{T}}_{k-1})). \qquad (5)$$

Now, for any $S \subset U$ and all $\boldsymbol{\mathcal{T}} \in U$, let us define $\Pi_S : U \mapsto 2^S$, where 2^S denotes the power set of S, as $\Pi_S(\boldsymbol{\mathcal{T}}) =$

$\arg\min_{\boldsymbol{\mathcal{X}}\in S}\|\boldsymbol{\mathcal{X}}-\boldsymbol{\mathcal{T}}\|_F$. Clearly, if S is a closed nonempty convex set, $\Pi_S(\boldsymbol{\mathcal{T}})$ contains exactly one element: the projection of $\boldsymbol{\mathcal{T}}$ onto S. If S is not convex but is closed and nonempty, then $\Pi_S(\boldsymbol{\mathcal{T}})$ is still nonempty (by the extreme value theorem, coercivity and continuity of $\|\boldsymbol{\mathcal{X}}-\boldsymbol{\mathcal{T}}\|_F$), but might contain multiple elements. Relying on the definition of Π_S, the next result shows how a minimizer of J_k over $L_{\mathbf{r}}$ can be obtained from $\boldsymbol{\mathcal{T}}^\star$.

Proposition 3.1. Let $S\subset U$ be a closed nonempty set. Then, $\Pi_S(\boldsymbol{\mathcal{T}}^\star)$ is the set of minimizers of $J_k(\boldsymbol{\mathcal{T}})$ over S, where $\boldsymbol{\mathcal{T}}^\star$ is given by (5).

Proof. Since S is closed and nonempty, J_k is continuous and $J_k(\boldsymbol{\mathcal{T}})\to\infty$ for $\|\boldsymbol{\mathcal{T}}\|_F\to\infty$, J_k admits at least one minimum in S. Also, for any $\boldsymbol{\mathcal{T}}\in S$, we can write $\boldsymbol{\mathcal{T}}=\boldsymbol{\mathcal{T}}^\star+\boldsymbol{\mathcal{Z}}$ for some $\boldsymbol{\mathcal{Z}}\in U$ and then rewrite J_k as

$$\begin{aligned}J_k(\boldsymbol{\mathcal{T}}^\star+\boldsymbol{\mathcal{Z}}) = {}& J_k(\boldsymbol{\mathcal{T}}^\star)+\|\boldsymbol{\mathcal{Z}}\|_F^2+2\langle\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{T}}^\star-\boldsymbol{\mathcal{T}}_{k-1}\rangle\\ &-2\mu_k\langle\mathscr{A}(\boldsymbol{\mathcal{Z}}),\mathbf{y}-\mathscr{A}(\boldsymbol{\mathcal{T}}_{k-1})\rangle\\ = {}& J_k(\boldsymbol{\mathcal{T}}^\star)+\|\boldsymbol{\mathcal{Z}}\|_F^2+2\langle\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{T}}^\star-\boldsymbol{\mathcal{T}}_{k-1}\rangle+2\mu_k\langle\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{G}}_{k-1}\rangle\\ = {}& J_k(\boldsymbol{\mathcal{T}}^\star)+\|\boldsymbol{\mathcal{Z}}\|_F^2,\end{aligned}$$

where the last equality follows directly from (5). Hence, as $\boldsymbol{\mathcal{Z}}=\boldsymbol{\mathcal{T}}-\boldsymbol{\mathcal{T}}^\star$, we have the equivalence

$$\arg\min_{\boldsymbol{\mathcal{T}}\in S}J_k(\boldsymbol{\mathcal{T}})=\arg\min_{\boldsymbol{\mathcal{T}}\in S}\|\boldsymbol{\mathcal{T}}-\boldsymbol{\mathcal{T}}^\star\|_F^2=\Pi_S(\boldsymbol{\mathcal{T}}^\star).$$

□

Since $L_{\mathbf{r}}$ is closed and nonempty, we have from Proposition 3.1 that, if there is *some* μ_k such that $J_k(\boldsymbol{\mathcal{T}})$ majorizes $\mu_k J(\boldsymbol{\mathcal{T}})$ and for which there exists $\boldsymbol{\mathcal{T}}_k\in\Pi_{L_{\mathbf{r}}}(\boldsymbol{\mathcal{T}}_{k-1}+\mu_k\boldsymbol{\mathcal{G}}_{k-1})$ satisfying $\boldsymbol{\mathcal{T}}_k\neq\boldsymbol{\mathcal{T}}_{k-1}$, then $J(\boldsymbol{\mathcal{T}}_k)<J(\boldsymbol{\mathcal{T}}_{k-1})$. It turns out, however, that the requirement of having a $J_k(\boldsymbol{\mathcal{T}})$ that majorizes $\mu_k J(\boldsymbol{\mathcal{T}})$ for all $\boldsymbol{\mathcal{T}}$ can be relaxed to improve convergence speed, as discussed in the next section.

It should be noted that it is hard to ensure in practice that $\boldsymbol{\mathcal{T}}_k$ is indeed a minimum of J_k over $L_{\mathbf{r}}$, because projecting onto $L_{\mathbf{r}}$ is not an easy task. To avoid an excessive computational cost per iteration, TIHT employs the quasi-optimal projection $\mathscr{H}_{\mathbf{r}}$, and thus $\boldsymbol{\mathcal{T}}_k$ is only *close to* a minimum. Nevertheless, as observed by [13], practical experience suggests that this suboptimality does not preclude TIHT from converging. Yet, a more rigorous analysis taking it into account remains as a topic for future investigation.

Remark 3.2. Interestingly, Proposition 3.1 is quite general, being valid for *any* closed nonempty subset S. Thus, the above reasoning clearly holds for other formulations as, *e.g.*, one based on the rank definition which applies to the tensor train model (see [8] and references therein). More generally, it can be extended to any problem of the form (1) in a finite-dimensional Hilbert space, as long as the feasible set is closed and nonempty.

4. IMPROVED STEP SELECTION STRATEGY

In spite of the successful recovery results shown in [13], the suitability of the step size formula (3) for achieving actual decrease of J is not discussed. This formula provides the optimal gradient descent step for *unconstrained* minimization (*i.e.*, over U). However, when minimizing over $L_{\mathbf{r}}$ with the scheme (2), its optimality is lost. Equally importantly from a practical standpoint, the behavior of the resulting algorithm is not satisfactory, because it converges quite slowly.

In the previous section, we have used the inequality $\mu_k J(\boldsymbol{\mathcal{T}}_k)<J_k(\boldsymbol{\mathcal{T}}_k)$ to derive $J(\boldsymbol{\mathcal{T}}_k)<J(\boldsymbol{\mathcal{T}}_{k-1})$. Thus, it suffices to guarantee that this inequality holds at $\boldsymbol{\mathcal{T}}_k$, instead of requiring that $J_k(\boldsymbol{\mathcal{T}})$ majorizes $\mu_k J(\boldsymbol{\mathcal{T}})$ at all $\boldsymbol{\mathcal{T}}$. To this end, one can check whether μ_k satisfies

$$\mu_k<\omega(\mu_k)=\frac{\|\boldsymbol{\mathcal{T}}_k-\boldsymbol{\mathcal{T}}_{k-1}\|_F^2}{\|\mathscr{A}\left(\boldsymbol{\mathcal{T}}_k-\boldsymbol{\mathcal{T}}_{k-1}\right)\|_2^2},\tag{6}$$

because such inequality, together with (4), implies $\mu_k J(\boldsymbol{\mathcal{T}}_k)<J_k(\boldsymbol{\mathcal{T}}_k)$. Note that the notation $\omega(\mu_k)$ emphasizes that the bound for μ_k depends on μ_k itself.

The condition (6) is an extension of that proposed in [14] for NIHT, which ensures cost function decrease when an optimal step cannot be computed with a simple formula. In that situation, NIHT only accepts a candidate step size if it satisfies a condition analogous to (6); otherwise, it is reduced until that condition is fulfilled. In the case of TIHT, (6) was not violated by step sizes computed with (3) during our practical experiments. However, (3) often yields $\mu_k\ll\omega(\mu_k)$, while empirical evidence suggests that the optimal step lies usually closer to its bound. This slows down the convergence of the algorithm. Note that, since any μ_k satisfying $\mu_k<\|\mathscr{A}\|^{-2}$ is majorized by (3), this observation also justifies the use of the more relaxed condition (6).

To circumvent this problem, we propose a modified algorithm, named improved-step-selection TIHT (ISS-TIHT), in which a heuristic step selection routine is added. The idea behind this routine is simple: given a fixed α such that $0\ll\alpha<1$, one checks whether the candidate μ_k satisfies

$$\alpha\omega(\mu_k)\leq\mu_k<\omega(\mu_k),\tag{7}$$

keeping its associated estimate $\boldsymbol{\mathcal{T}}_k$ when it does. Otherwise, we simply set $\mu_k=\beta\omega(\mu_k)$ for some $\beta\in(\alpha,1)$, compute a new $\boldsymbol{\mathcal{T}}_k$ and repeat the process. We employ (3) as the starting candidate μ_k, which is reasonable since it satisfies (7) at some iterations. As there is no guarantee of finding a step fulfilling (7) with this procedure, we establish a maximum number of trials L, after which we keep the biggest generated step size satisfying the upper bound of (7). If none of them does, we take the smallest candidate step and proceed as in the NIHT [14], reducing it via division by a factor $\kappa>1$ until the upper bound is verified. A pseudocode describing this scheme is shown in Algorithm 1, where we denote the candidate values of μ_k and $\boldsymbol{\mathcal{T}}_k$ by $\mu_{k,l}$ and $\boldsymbol{\mathcal{T}}_{k,l}$, respectively, for $l=1,\ldots,L$.

Algorithm 1 ISS-TIHT

for $k = 1, 2, \ldots, K$ **do**
  $\mathcal{G}_{k-1} = -\mathscr{A}^*(\mathbf{y} - \mathscr{A}(\mathcal{T}_{k-1}))$
  $\mu_{k,1} = \|\mathcal{G}_{k-1}\|_F^2 \, \|\mathscr{A}(\mathcal{G}_{k-1})\|_2^{-2}$
  for $l = 1, \ldots, L$ **do**
    $\mathcal{T}_{k,l} = \mathscr{H}_{\mathbf{r}}(\mathcal{T}_{k-1} - \mu_{k,l}\mathcal{G}_{k-1})$
    if $\alpha\omega(\mu_{k,l}) \leq \mu_{k,l} \leq \omega(\mu_{k,l})$ **then**
      select $\mu_k = \mu_{k,l}$, $\mathcal{T}_k = \mathcal{T}_{k,l}$
      break
    end if
    $\mu_{k,l+1} = \beta\omega(\mu_{k,l})$
  end for
  if no $\mu_{k,l}$ was selected **then**
    if $\exists\, l$ such that $\mu_{k,l} < \omega_k(\mu_{k,l})$ **then**
      $l^\star = \arg\max_l \mu_{k,l}$ subject to $\mu_{k,l} < \omega_k(\mu_k)$
    else
      $l^\star = \arg\min_l \mu_{k,l}$
      while $\mu_{k,l^\star} \geq \omega(\mu_{k,l^\star})$ **do**
        $\mu_{k,l^\star} = \mu_{k,l^\star}/\kappa$
      end while
    end if
    select $\mu_k = \mu_{k,l^\star}$, $\mathcal{T}_k = \mathcal{T}_{k,l^\star}$
  end if
end for

We point out that the idea of choosing a new candidate for the step size as $\beta\omega(\mu_k)$ is suggested in [14] with $\beta = 1$, but only in the case that the current candidate violates the upper bound $\omega(\mu_k)$. In other words, a candidate step size is never *increased* in NIHT; rather, it is only *shrunk* if the upper bound $\omega(\mu_k)$ is not met. In the case of TIHT, enforcing also the lower bound of (7) substantially accelerates convergence.

5. SIMULATION RESULTS

We now evaluate ISS-TIHT in two simulation scenarios. First, a LRTR setting with an unconstrained operator and a synthetic data tensor is considered. To this end, we randomly generate $\mathscr{A}$ and $\mathcal{T} \in \mathbb{R}^{N\times N\times N}$ and apply four algorithms to recover $\mathcal{T}$ from $\mathbf{y} = \mathscr{A}(\mathcal{T}) = \mathbf{A}\,\mathrm{vec}(\mathcal{T})$, where $\mathbf{A} \in \mathbb{R}^{M\times N^3}$, with $M = \rho N^3$. The data tensor is given by $\mathcal{T} = \mathcal{T}_0 + 10^{-5}\mathcal{N}$, where $\mathcal{T}_0$ is a low-rank tensor generated via $\mathcal{T}_0 = \mathcal{S} \times_1 \mathbf{V}^{(1)} \times_2 \mathbf{V}^{(2)} \times_3 \mathbf{V}^{(3)}$, with $\mathcal{S} \in \mathbb{R}^{R\times R\times R}$ and $\mathbf{V}^{(p)} \in \mathbb{R}^{N\times R}$. All $\mathbf{A}$, $\mathcal{N}$, $\mathcal{S}$ and $\mathbf{V}^{(p)}$ have standard Gaussian i.i.d. elements, and we normalize $\mathcal{T}_0$ and $\mathcal{N}$ so that $\|\mathcal{T}_0\|_F = \|\mathcal{N}\|_F = 1$. The evaluated algorithms are TIHT, ISS-TIHT, an alternating direction method of multipliers (ADMM) scheme based on that of [10, Sec. 4.4], which minimizes a weighted sum of NNs of matrix unfoldings, and a generalized alternating least-squares (GALS) scheme, which estimates the components of a low-m-rank Tucker model of $\mathcal{T}$ by minimizing J with respect to them in an alternating fashion. We empirically set the regularization and penalty parameters of ADMM respectively as $\lambda = 10^{-2}$ and $\eta = 2$, aiming at maximizing estimation precision, and use weights $\gamma_1 = \gamma_2 = \gamma_3 = 1/3$. Regarding ISS-TIHT, we use $\alpha = 0.5$, $\beta = 0.7$, $\kappa = 1.2$ and $L = 5$. We fix $R = 5$, $N = 20$, $\rho = 0.15$ and let all the algorithms run for $K = 60$ iterations, measuring at each iteration k the normalized square error $\mathrm{NSE}_k = \|\mathcal{T} - \mathcal{T}_k\|_F^2/\|\mathcal{T}\|_F^2$. TIHT, ISS-TIHT and GALS are run with $R_1 = R_2 = R_3 = R$. This procedure is repeated for 30 joint realizations of $\mathbf{A}$, $\mathcal{T}_0$ and $\mathcal{N}$, and the average NSE_k is computed for each k and each algorithm, yielding the corresponding normalized mean-square error NMSE_k displayed in Fig. 1. For reference, we plot also the NMSE of the (R, R, R)-truncated HOSVD of $\mathcal{T}$. Table 1 reports the average computing time per iteration measured in a Intel Xeon ES-2630v2 2.60 GHz. The results show that ISS-TIHT converges much faster than TIHT. GALS converges even faster, but at the expense of a much higher computational cost. Since ADMM performs poorly with M measurements, we also evaluate it using the same procedure but with $\mathbf{A}$ providing $2M$ and $3M$ measurements. As the curves show, only with $3M$ measurements ADMM attains low error (at a quite high cost), but is still outperformed by ISS-TIHT and GALS.

In the second scenario, we consider a TC setting. The data tensor $\mathcal{T} \in \mathbb{R}^{128\times 128\times 128}$ now contains the brain MRI data used in [17]. The evaluated algorithms are the ADMM scheme of [10, Sec. 4.4], TIHT and ISS-TIHT. GALS and the approach of [17] are not included since the first is too costly for this setting, while the latter does not apply to TC. The ADMM algorithm is employed with the observations as constraints ($\lambda \to 0$), as proposed in [10] for the noiseless

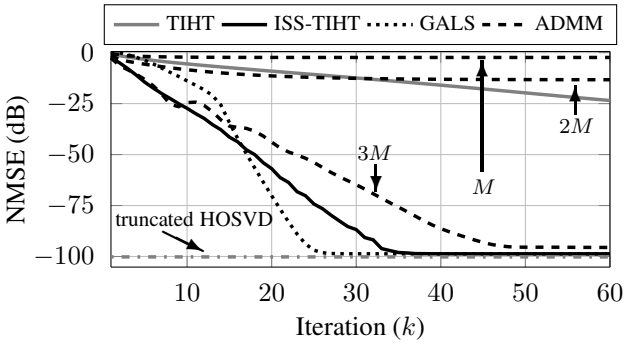

Fig. 1. Average NMSE measured in a LRTR setting with unconstrained operator (scenario 1).

Algorithm	Scenario 1 (LRTR)	Scenario 2 (TC)
TIHT	1.78×10^{-2}	1.65×10^{-1}
ISS-TIHT	2.86×10^{-2}	3.41×10^{-1}
GALS	4.13×10^{-1}	—
ADMM (M)	2.61×10^{-1}	3.96×10^{0}
ADMM ($2M$)	6.67×10^{-1}	—
ADMM ($3M$)	1.28×10^{0}	—

Table 1. Average time measured per iteration (in seconds).

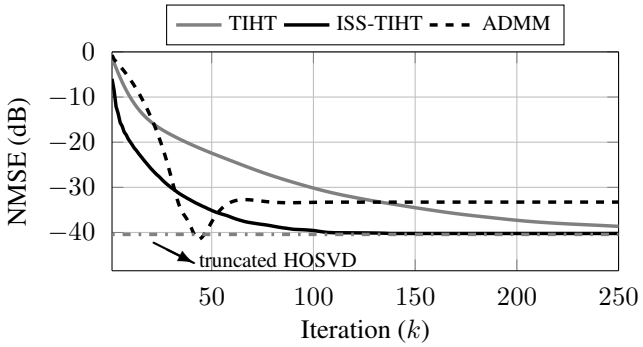

Fig. 2. Average NMSE measured in a TC setting (scenario 2).

case. We set (again, empirically) $\eta = 0.04$, and choose $\gamma_1 = \gamma_2 = \gamma_3 = 1/3$. For ISS-TIHT, we use $\alpha = 0.5$, $\beta = 0.7$, $\kappa = 1.2$ and $L = 5$. TIHT and ISS-TIHT are run with $R_1 = R_2 = R_3 = R = 20$. To generate $\mathbf{A}$ and $\mathbf{y}$, we randomly choose a subset of the indices of $\boldsymbol{\mathcal{T}}$ of cardinality $M = \rho 128^3$, with $\rho = 0.15$, filling $\mathbf{y}$ with the corresponding elements. This process was repeated for 30 realizations of $\mathbf{A}$. The results are shown in Table 1 and Fig. 2. Again, we plot the NSE of the (R, R, R)-truncated HOSVD of $\boldsymbol{\mathcal{T}}$, which is an approximate lower bound for the NMSE_k of both TIHT and ISS-TIHT. One can see that, thanks to the introduced step selection heuristic, ISS-TIHT outperforms both TIHT and ADMM.

Finally, we point out that, in all our experiments, ISS-TIHT always found at least a candidate step size satisfying the upper bound of (7). Also, the behavior of the algorithm was not observed to be too sensitive to the choice of α and β.

6. CONCLUSION

We have studied the TIHT algorithm by relying on the optimization strategy which underlies the NIHT algorithm. This offers an insightful interpretation of its iterates, whose cost function values are monotonically decreasing under a certain upper bound on the step size, assuming the best low-rank approximation calculated at each iteration is exact. Then, we have proposed the ISS-TIHT algorithm, which includes a subroutine that attempts to find a step within a constant factor of its bound. Our simulation results show that this simple heuristic leads to a remarkable acceleration of convergence.

REFERENCES

[1] B. Recht et al., "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM review*, vol. 52, no. 3, pp. 471–501, 2010.

[2] J. Liu et al., "Tensor completion for estimating missing values in visual data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 208–220, 2013.

[3] N. Kreimer and M. D. Sacchi, "A tensor higher-order singular value decomposition for prestack seismic data noise reduction and interpolation," *Geophysics*, vol. 77, no. 3, pp. V113–V122, 2012.

[4] M. Signoretto et al., "Tensor versus matrix completion: A comparison with application to spectral data," *IEEE Signal Process. Lett.*, vol. 18, no. 7, pp. 403–406, 2011.

[5] B. Romera-Paredes et al., "Multilinear multitask learning," in *Proc. 30th Int. Conf. Mach. Learning*, 2013, pp. 1444–1452.

[6] C. J. Hillar and L.-H. Lim, "Most tensor problems are NP-hard," *J. ACM*, vol. 60, no. 6, pp. 45:1–45:39, 2013.

[7] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM review*, vol. 51, no. 3, pp. 455–500, 2009.

[8] L. Grasedyck et al., "A literature survey of low-rank tensor approximation techniques," *GAMM-Mitteilungen*, vol. 36, no. 1, pp. 53–78, 2013.

[9] S. Oymak et al., "Simultaneously structured models with application to sparse and low-rank matrices," *arXiv preprint arXiv:1212.3753*, 2012.

[10] R. Tomioka et al., "Estimation of low-rank tensors via convex optimization," *arXiv:1010.0789*, 2010.

[11] C. Mu, B. Huang, J. Wright, and D. Goldfarb, "Square deal: Lower bounds and improved relaxations for tensor recovery," in *Proc. 31st Int. Conf. Mach. Learning*, 2014, pp. 73–81.

[12] A. Uschmajew, "A new convergence proof for the high-order power method and generalizations," *arXiv:1407.4586*, 2014, to appear in Pac. J. Optim.

[13] H. Rauhut et al., "Low rank tensor recovery via iterative hard thresholding," in *Proc. 10th Int. Conf. Sampling Theory Applicat.*, 2013.

[14] T. Blumensath and M. E. Davies, "Normalized iterative hard thresholding: Guaranteed stability and performance," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 2, pp. 298–309, 2010.

[15] G. Allaire, *Numerical Analysis and Optimization*, Oxford University Press, 2007.

[16] M. Ishteva et al., "Best low multilinear rank approximation of higher-order tensors, based on the Riemannian trust-region scheme," *SIAM J. Matrix Anal. Applicat.*, vol. 32, no. 1, pp. 115–135, 2011.

[17] C. F. Caiafa and A. Cichocki, "Stable, robust, and super fast reconstruction of tensors using multi-way projections," *IEEE Trans. Signal Process.*, vol. 63, no. 3, pp. 780–793, 2015.